REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-06-2006	Final	Dec 2001 - Mar 2006

4. TITLE AND SUBTITLE

An Investigation of the Reliability Of Knowledge Measures Through Relational Mapping in Joint Military Environments

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-02-1-0179

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John J. Lee, Gregory K. W. K. Chung, William L. Bewley, Alicia M. Cheak, and Karen Ellis

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UCLA CSE/CRESST
300 Charles E. Young Dr. North
300 GSE&IS/Mailbox 951522
Los Angeles, CA 90095-1522

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report describes research conducted by the UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST) on the effectiveness of online knowledge mapping as a method to assess high-level understanding of specific military domains and tasks. Using the Human Performance Knowledge Mapping Tool (HPKMT), CRESST assessed individual trainee knowledge and then examined the psychometric properties of knowledge mapping scores to evaluate the suitability of knowledge mapping as an assessment of trainees' understanding of joint mission-essential tasks. Analyses of scoring techniques yielded important information about the quality of the knowledge maps, and the assessments provided valuable information regarding student understanding of course content. The student maps were analyzed using three methods: automated criterion-based (expert) assessment, propositional analysis, and structural mapping analysis.

15. SUBJECT TERMS

online knowledge mapping, automated scoring

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(Include area code)*
Unclassified	Unclassified	Unclassified			

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

An Investigation of the Reliability of Knowledge Measures Through Relational Mapping in Joint Military Environments

Deliverable – June 2006

Knowledge, Models and Tools to Improve
the Effectiveness of Naval Distance Learning

Eva L. Baker
CRESST/University of California, Los Angeles

Office of Naval Research
Award # N00014-02-1-0179

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Office of Naval Research, Award Number #N00014-02-1-0179, as administered by the Office of Naval Research.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the Office of Naval Research, or the U.S. Department of Defense.

TABLE OF CONTENTS

AN INVESTIGATION OF THE RELIABILITY OF KNOWLEDGE MEASURES THROUGH RELATIONAL MAPPING IN JOINT MILITARY ENVIRONMENTS

John J. Lee, Gregory K. W. K. Chung, William L. Bewley, Alicia M. Cheak, and Karen Ellis

CRESST/University of California, Los Angeles

Executive Summary

In response to a growing need for distance learning that is provided just-in-time, this study looked at the reliability of knowledge mapping measures developed at CRESST, with the Human Performance Knowledge Mapping Tool (HPKMT). The HPKMT can be used for assessment purposes and has the capability to automatically score knowledge maps against expert maps. The HPKMT has been developed over many years and has been used in a myriad of educational contexts. The purpose of this study was to determine the reliability of these types of measures in a joint military environment.

Twenty-nine all-male military personnel from the Joint Special Operations University in Hurlburt, Florida, participated in this study. They were mostly from the Army, some from the Air Force, and one from the Navy and one was a civilian. They were fairly evenly split between enlisted members and officers. Students were asked to create three knowledge maps for three content areas: Air Tasking Order (ATO) cycle, Joint Task Force Structure and Function (JTF), and Joint Special Operation Task Force Structure (JSOTF). Due to scheduling and administrative constraints, insufficient numbers completed the JSOTF task so it was dropped from the analysis.

Because an insufficient number of participants was provided, we were unable to complete the generalizability analysis, but analyses of scoring techniques yielded important information about the quality of the knowledge maps, and the assessments provided valuable information regarding student understanding of JSOU course content.

Expert maps were elicited from four experts and used as criterion maps for scoring. The student maps were analyzed using three methods: automated criterion-based (expert) assessment, propositional analysis, and structural mapping analysis. The criterion-based assessment showed significantly lower scores for the students as compared to experts for both tasks. The propositional analysis found that the expert and student use of terms and links were fairly proportional, with some exceptions. There were items where experts have different frequencies than those of students. For example, *apportionment* was the highest source term on the ATO cycle task for experts (at 7.8%) while for students it was much lower (1.5%). Experts relied on more functional links for ATO, e.g. *supports (25%), input (21%), output (19%),* and students use more relational links for ATO, e.g. *leads to (32%), followed by (14%), supports (14%).* The final analysis looked at the structural nature of the maps or how concepts are connected to each other. An analysis of the structure revealed differences between expert maps and student maps, and differences among students' maps relative to structural complexity. In general, the expert maps had more terms; variable use of source, sinks, and carriers; numerous clusters; and high reachability. Additionally, a comparison of a sample of student maps revealed similar patterns, with more sophisticated maps containing a higher number of terms, links, and clusters as well as level of reachability.

In addition to these research results, we were able to create a standalone version of the mapper that has been used in subsequent studies for the military.

AN INVESTIGATION OF THE RELIABILITY OF KNOWLEDGE MEASURES THROUGH RELATIONAL MAPPING IN JOINT MILITARY ENVIRONMENTS

John J. Lee, Gregory K. W. K. Chung, William L. Bewley, Alicia M. Cheak, and Karen Ellis

CRESST/University of California, Los Angeles

Introduction

The armed services are turning increasingly to advanced distributed learning (ADL) systems to deliver training and education solutions on a global scale. A common expectation for ADL systems is the delivery of quality training—to the right people, at the right time, and at the right place—to support operational readiness and personal excellence (e.g., Air Force Institute for Advanced Distributed Learning, 2001; Department of Defense, 1999; Director of Naval Training [N7], 1998; U.S. Army Training and Doctrine Command, 1999). The development of the technical infrastructure and standards is currently underway (e.g., Advanced Distributed Learning [ADL], 2003a) as well as guidelines for effective ADL implementation (ADL, 2003b).

While much of the focus has been on the delivery of instructional content to trainees via ADL, an important complement to instruction is assessment. Effective training and education are facilitated by the capability to measure the degree to which trainees have attained the intended outcomes of training and instruction. Assessment capability can also provide information on, for example, estimates of what trainees know prior to training, how much they have learned from training, how well they may perform in a future situation, and whether to recommend remediation content to bolster a trainee's knowledge. Finally, just as with instructional components, ADL-based assessment must be sensitive to the underlying drivers behind the ADL initiative, such as cost-effective delivery of assessments, an uncertain budget environment, decreased personnel strengths, increased deployments, and rapidly changing missions.

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has refined the Human Performance Knowledge Mapping Tool (HPKMT) to support rapid, automated, and cost-effective assessment of domain knowledge. The system was developed with support from the Office of Naval Research Capable Manpower Future Naval Capability initiative. The tool is designed to assess a trainee's understanding of a content domain via graphical representation. Trainees are required to express their understanding of a content area by creating knowledge maps. Knowledge maps are network representations, where nodes represent concepts and links represent the relationship between two concepts.

The basic measurement approach has been tested in numerous educational settings outside the military context. The focus of the proposed work was on gathering evidence of the effectiveness of online knowledge mapping as a method to assess high-level understanding of specific military domains and tasks. Thus, we proposed to use our HPKMT to assess individual trainee knowledge (i.e., a trainee maps his or her understanding of the domain using our online knowledge mapping tool), and then examined the psychometric properties of knowledge mapping scores to evaluate the suitability of knowledge mapping as an assessment of trainees' understanding of joint mission-essential tasks.

Reliability and Validity of Knowledge Maps as an Assessment

A presumed critical capability of an assessment in a distributed learning setting is automated scoring. A critical validity issue of an assessment is the scoring, regardless of automated capability. In this section we briefly describe the different types of scoring and provide examples of their use. For in-depth reviews of assessment issues related to knowledge maps, see Ruiz-Primo and Shavelson (1996).

In general, scoring knowledge maps can be referent-based or referent-free. Referent-based methods compare a student's map against a referent map (e.g., an expert's map or other gold standard). Referent-free methods evaluate the student's map against a rubric or with other criteria (e.g., judging the quality of the propositions [node-link-node relation], or counting the number of concepts in the map). In either case, different scoring approaches use to different degrees the configural and semantic properties of the network. Table 1 summarizes scoring methods.

Table 1

Simplified Summary of Knowledge Mapping Scoring Methods

	Configural	Semantic
Referent-free	Explicitly scores a map or elements of a map on its structural aspect (e.g., considering degree of hierarchical organization). Example application: Novak and Gowin (1984).	Explicitly scores a map or elements of a map on its semantic aspect (e.g., scoring quality of propositions). Example applications: Osmundson, Chung, Herl, and Klein (1999). Shavelson (Ruiz-Primo, Schultz, Li, & Shavelson, 2001)
Referent-based	Compares the network structure of a student's map and the referent map. Does not take into account the meaning of the relationships. Example application: Herl, Baker, and Niemi (1996).	Compares the semantic structure of a student's map and the referent map (e.g., proposition-by-proposition comparison between a student's map and an expert's map). Ignores the configural aspects of the network. Example applications: Herl et al. (1996). Osmundson et al. (1999).

Referent-Free Scoring Methods

The scoring procedure specified by Novak and Gowin (1984) is one of the earliest and most commonly used methods of scoring knowledge maps. Their method considers hierarchy as an important component of the scoring, as well as propositions, cross-links, and examples. In terms of hierarchy, credit is given for each hierarchical level showing subordinate concepts at a lower level as more specific than their parent concepts. Each valid and meaningful proposition is also credited, as are examples and cross-links. Cross-links are links between different hierarchical levels. Novak and Gowin's scoring scheme is weighted heavily towards the hierarchical structure of the map. The theoretical rationale for this scoring scheme is Ausubel's theory of learning, particularly the ideas of subsumption (new ideas can be subsumed under more general concepts) and progressive differentiation (as learning occurs, there is more differentiation among the concepts, which is shown by the inclusion of more propositions and cross-links).

Evidence from several studies suggests that Novak and Gowin's (1984) scoring scheme can differentiate between high- and low-knowledge students in biology (Markham, Mintzes, & Jones, 1994) and between first-year and advanced pediatric residents studying seizures (West, Pomeroy, Park, Gerstenberger, & Sandoval, 2000). This scoring scheme also appears to be sensitive to learning, as student map scores increased over instruction (Pearsall, Skipper, & Mintzes, 1997; West et al., 2000).

A second scoring scheme that is commonly used considers only the propositions contained in the map and not the configural aspects. This method is to rate the quality of the propositions in the map. Each proposition is evaluated in terms of its accuracy. For example, Ruiz-Primo and colleagues used a proposition accuracy score as one measure of the quality of students' knowledge maps (Ruiz-Primo, Schultz, Li, & Shavelson, 1997; Ruiz-Primo et al., 2001). Each proposition in a student's map was scored on a 5-point scale, ranging from 0 (invalid/inaccurate) to 4 (complete and correct and showing a deep understanding of the relation between two concepts). Ruiz-Primo and colleagues found that students' proposition accuracy scores differentiated high-knowledge students from low-knowledge students (e.g., Ruiz-Primo et al., 1997) and students' map scores were moderately correlated (r between .40 to .50) with other measures of content knowledge formats (e.g., essays, multiple choice tests). Similar relationships have been found between knowledge map proposition accuracy scores and classroom end-of-unit tests and standardized tests of reading, math, and science (Rice, Ryan, & Samson, 1998), and between knowledge maps and physics problem solving (Austin & Shore, 1995).

Referent-Based Scoring Methods

Referent-based scoring methods compare a student's map against a criterion map. Example referents include an expert's map, a composite map of experts, or the instructor's map. The essential measure is the number of propositions in the student map that are also in the referent map. Several studies have investigated the technical properties of this approach. For example, Ruiz-Primo et al. (2001), in addition to using proposition accuracy scores, also scored students' maps against an expert's map. The correlation between the proposition accuracy score and expert-based score was sufficiently high for Ruiz-Primo et al. to conclude that an expert-based method was the most efficient scoring method (i.e., in terms of scoring time and reliability of scores). Similar results were found by Osmundson et al. (1999) and Chung, Harmon, and Baker (2001).

The findings of Ruiz-Primo et al. (2001) are consistent with earlier work by Herl (1995), Herl et al. (1996), and Osmundson et al. (1999). In general, scoring student knowledge maps using expert-based referents has been found to discriminate between experts and novices (Herl, 1995; Herl et al., 1996), discriminate between different levels of student performance (Herl, 1995; Herl et al., 1996), relate moderately to external measures (Aguirre-Muñoz, 2000; Herl, 1995; Herl et al., 1996; Klein, Chung, Osmundson, Herl, & O'Neil, 2002; Lee, 1999; Osmundson et al., 1999), detect changes in learning (Chung et al., 2001; Osmundson et al., 1999; Schacter, Herl, Chung, Dennis, & O'Neil, 1999), and be sensitive to language proficiency (Aguirre-Munoz, 2000; Lee, 2000).

The final type of scoring is to simply compare the network topology of a student's map and the referent map. Herl et al. (1996) investigated the utility of this approach and found high correlations between scores based on a comparison of the network topology and scores based on the overlap of propositions between the student and expert map.

Generalizability of Knowledge Map Scores

To date, we could find only a few studies that have examined the generalizability of knowledge map scores (Cawley, Zimmaro, Van Meter, & Theodorou, 1999; McClure, Sonak, & Suen, 1999; Ruiz-Primo et al., 1997, 2001; Zimmaro, Zappe, Parkes, & Suen, 1999). In all cases, these studies used raters to score knowledge maps and thus raters were included as a facet. In two G studies (person × rater × task) conducted by Ruiz-Primo et al. (1997), proposition accuracy scores were found to have negligible rater effects. The largest variance component was due to persons, followed by the person × task interaction. The absolute and relative g-coefficients for these studies were in the high .80s. In a second series of G studies (person × rater) conducted on three different scoring methods (proposition quality, expert-criterion, "salience"), Ruiz-Primo et al. (2001) again found negligible rater effects. The absolute and relative g-coefficients, regardless of scoring method, were extremely high (high .90s). Similar high g-coefficients were reported by Zimmaro et al.; however, not all generalizability studies have shown negligible rater effects (e.g., see Cawley et al., 1999; McClure et al., 1999).

Summary

A variety of approaches have been used to score knowledge maps, each with advantages and disadvantages. Overall, the cumulative findings reported across the various studies cited previously suggest that knowledge mapping is promising as a technique to measure students' knowledge of a domain. Knowledge map scores appear to differentiate between high- and low-knowledge students, to be sensitive to learning, to relate to other measures of performance, and to be sensitive to language proficiency.

From the perspective of DL, referent-based methods are the most suitable for automated scoring approaches. Referent-free scoring methods are less favorable for automated scoring because the approach usually attempts to measure quality. For example, Novak and Gowin's (1984) scoring technique requires evaluation of the map with respect to both structure and accuracy. Raters need to judge the degree of accuracy of links across hierarchical levels in the map. Similarly, while simpler, the proposition quality method requires raters to evaluate the accuracy of each proposition and assign a score. Automating the proposition accuracy technique is tractable when the set of concepts and links remain fixed, and when trainees are not permitted to generate their own idiosyncratic concepts and links.

Research Questions

While much research has been conducted on knowledge mapping in K-16 environments, there has been only limited work on examining the technical properties of knowledge maps in general, and virtually no work on examining the technical properties of online knowledge mapping for DL purposes, much less in a military context. Thus, we proposed to gather information on the reliability of online knowledge mapping in a military context. We proposed to address three questions:

How many criterion maps are necessary to achieve adequate reliability? Whereas prior generalizability (G) studies have included a rater facet to examine consistency in rendering scores, in automated scoring there is little question about the consistency in scoring. Rather, the issue (for the expert-criterion scoring method) is how many expert criterion maps are needed. This is an important practical question because gathering expert maps is straightforward and cost-effective compared to other methods of increasing reliability (e.g., increasing the number of tasks). In the past we have found high consistency of scores (e.g., see Herl, O'Neil, Chung, & Schacter, 1999; Klein et al., 2002) but little work has been done to gather information on the minimum number of criterion maps needed to achieve adequate reliability.

Does scoring stringency matter for reliability? With respect to stringency of scoring for expert-criterion maps, we have developed automated scoring methods for four levels of stringency: credit given only for an exact match between a proposition in the student map and a proposition in the expert map (stringent); credit given for matches between categories of links (ignoring the link direction); credit given for matches between the link direction (ignoring the link term); and credit given for matches between the link connection (ignoring the direction of the link and the link term). As the stringency decreases, student map scores tend to increase; what is unknown is whether stringency matters in terms of reliability and if so, which level of stringency results in the most reliable score.

How many mapping tasks are necessary to achieve adequate reliability? A common finding in performance assessments is the person × task interaction. That is, people perform differently on different tasks (Shavelson, Baxter, & Pine, 1991) and this effect has been observed for knowledge mapping tasks in particular (Ruiz-Primo et al., 1997). Unfortunately, there is no information on knowledge mapping in military contexts and there is no reason to expect that this effect will not be found. Thus, gathering information on the number of tasks required for adequate reliability is an important first step when applying knowledge mapping to a new domain.

Research Design

We proposed a series of generalizability and decision studies to address our research questions (Shavelson & Webb, 1991). The basic design was a person × rater × task design, where rater is the expert-criterion maps used to score students' maps. From this design, the following analyses can be conducted: Decision (D) studies to answer the question of how many expert criterion maps are needed for adequate reliability and number of tasks. Separate G studies will be conducted to answer scoring stringency questions.

Methodology

Design

Participants. Twenty-nine all-male military personnel (mean age = 41.56 years; mean number of years in military service = 18.76 years; mean number of years in Special Operations = 12.24 years) from the Joint Special Operations University participated in this study. In this sample, military branch division, rank, and educational background were mixed. Military branches include the Army (18 participants), Navy (1 participant), Air Force (5 participants) and other (1 participant). Rank was relatively evenly split between enlisted members (13 participants) and officers (10 participants), and only two participants were neither (1 civilian and 1 government service). Most of the participants had a college-level education, with 11 participants obtaining a 4-year college degree and 6 participants receiving a Master's, Doctoral, or Professional degree. Participants were students at the Joint Special Operations Task Force course at the Joint Special Operations University at Hurlburt, in Florida.

Classroom setting. Course curriculum was administered in the form of 30 lectures over five days and taught by several instructors and guest speakers. One of the instructors, having created an expert map with the Human Performance Knowledge Mapping Tool (HPKMT), was responsible for administering the mapping task. Due to scheduling restraints, participants received the mapping task at the end of Day 1 of the course. Participants were provided with an introduction to knowledge mapping, followed by a demo, and asked to create knowledge maps in three domains: Air Tasking Order (ATO), Joint Task Force Structure and Functions (JTF), and Joint Special Operation Task Force Structure (JSOTF).

Knowledge mapping system. The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has refined the Human Performance Knowledge Mapping Tool (HPKMT) to provide anytime, anywhere access capability for students and teachers. One feature of the HPKMT is its automated, referent-based scoring, which compares student maps against a criterion map. The essential measure is the number of propositions in the student map that are also present in the referent map.

Thus, we created a Web site that integrated the use of a relational database into the knowledge mapper. The main requirement for this site was to support the creation and maintenance and assessment of knowledge maps by students, teachers, and experts. The knowledge mapper was written in Java and was accessible from Internet Explorer browsers running on a Windows platform.

The user interface required only the use of the mouse. Concepts were added by dragging the concept icon to the map canvas and selecting a concept from a pop-up menu of available concepts. Links were created by connecting two concepts and then selecting the desired relationship label from a pop-up menu. The set of concepts and links was defined a priori, and no changes could be made directly to the terms and links in the knowledge mapper. *Figure 1* shows the main user interface of the knowledge mapper.

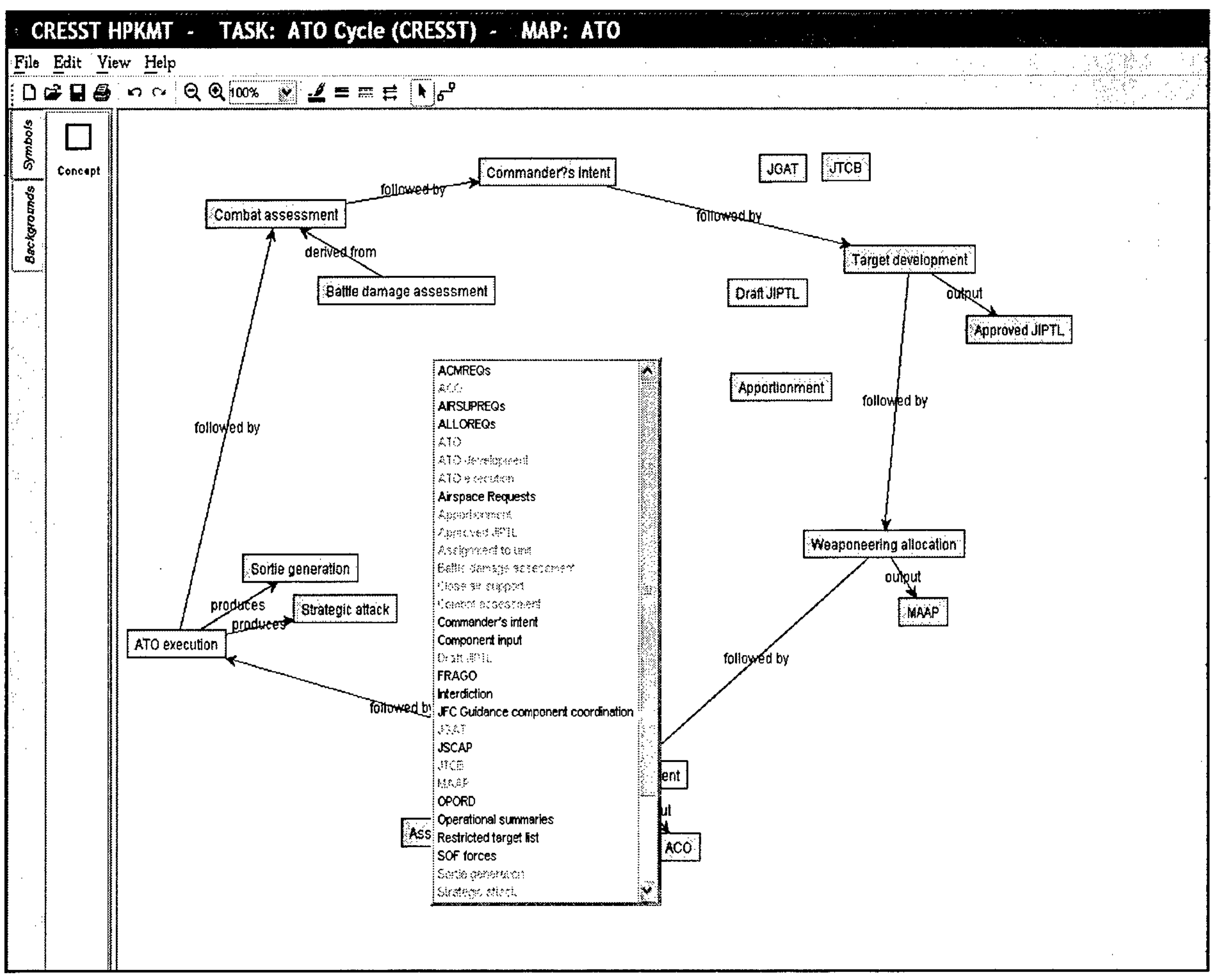

Figure 1. Example of a knowledge map (ATO Cycle).

Development of knowledge mapping terms and links. Four content experts (Department Head, Course Director, and two course instructors) deliberated over the learning objectives for the course and identified three key content areas: (1) ATO cycle, (2) JTF structures and functions, and (3) JSOTF structure. From this, experts worked together to generate a list of all possible terms relevant to the first domain, the ATO cycle. Each expert created a preliminary knowledge map using the list of concepts and generated linking terms to relate concepts to one another. The full set of concepts and links generated underwent review and modifications by the experts. The final knowledge mapping task for the ATO cycle contained 32 terms and 7 links, the JTF task had 36 terms and 14 links, and the JSOTF task had 51 terms and 19 links. See the appendix for a list of acronyms and abbreviations used in the study.

Experts each created final knowledge maps using the final list of terms and links. In all, 12 expert maps were generated, 4 for each domain. However, due to time constraints, as well as difficulties constraining the third task, JSOTF was removed. Table 2 summarizes the process of creating the list of concepts and links. Table 3 and Table 4 present the final list of concepts and links for ATO and Table 5 and Table 6 present the final list of concepts and links for the JTF task.

Table 2

Procedure Used to Generate Final Concepts and Links for Knowledge Mapping Task

Step	Procedure
1	Course experts reviewed relevant instructional materials.
2	Experts generated a list of all the possible terms relevant to each domain.
3	Preliminary set of terms and links reviewed and modified.
4	Final list of terms and links created.
5	Four experts each created a knowledge map using the final list of concepts and links.

Table 3

ATO Knowledge Map Concepts

ACMREQs	Battle damage assessment	Restricted target list
ACO	Close air support	SOF forces
AIRSUPREQs	Combat assessment	Sortie generation
ALLOREQs	Commander's intent	Strategic attack
ATO	Component input	Support
ATO development	Interdiction	Target development
ATO execution	JFC Guidance component coordination	TBMCS
Airspace Requests	JTCB	Weaponeering allocation
Apportionment	JGAT	Weaponization of targets
Approved JIPTL	MAAP	
Draft JIPTL	Operational summaries	
Assignment to unit		

Table 4

ATO Knowledge Map Links

Derived from
Followed by
Input
Leads to
Output
Produces
Supports

Table 5

JTF Knowledge Map Concepts

AFFOR	JCMOTF	NALE
AFSOA	JFACC	NAVFOR
AFSOC	JFC	NAVSOA
ARFOR	JFLCC	NAVSOC
ARSOA	JFMCC	NIST
ARSOC	JFSOCC	NSWTG
ARSOTF	JPOTF	NSWTU
BCD	JSOAC	OGA
C/JTF	JSOTF	RCC
Combatant CC	JTF	Service Components
CORP/MEF	MARFOR	SOCCE
Host Nation	MARLO	SOLE

Table 6

JTF Knowledge Map Links

ADCON
COCOM
OPCON
TACON
Allocated
Apportioned
Assigned
Attached
Coordinates
Liaison
Plans
Same as
Supports
Works for

The expert maps for ATO are shown in *Figure 2* and *Figure 3* and for JTF in *Figure 4* and *Figure 5*.

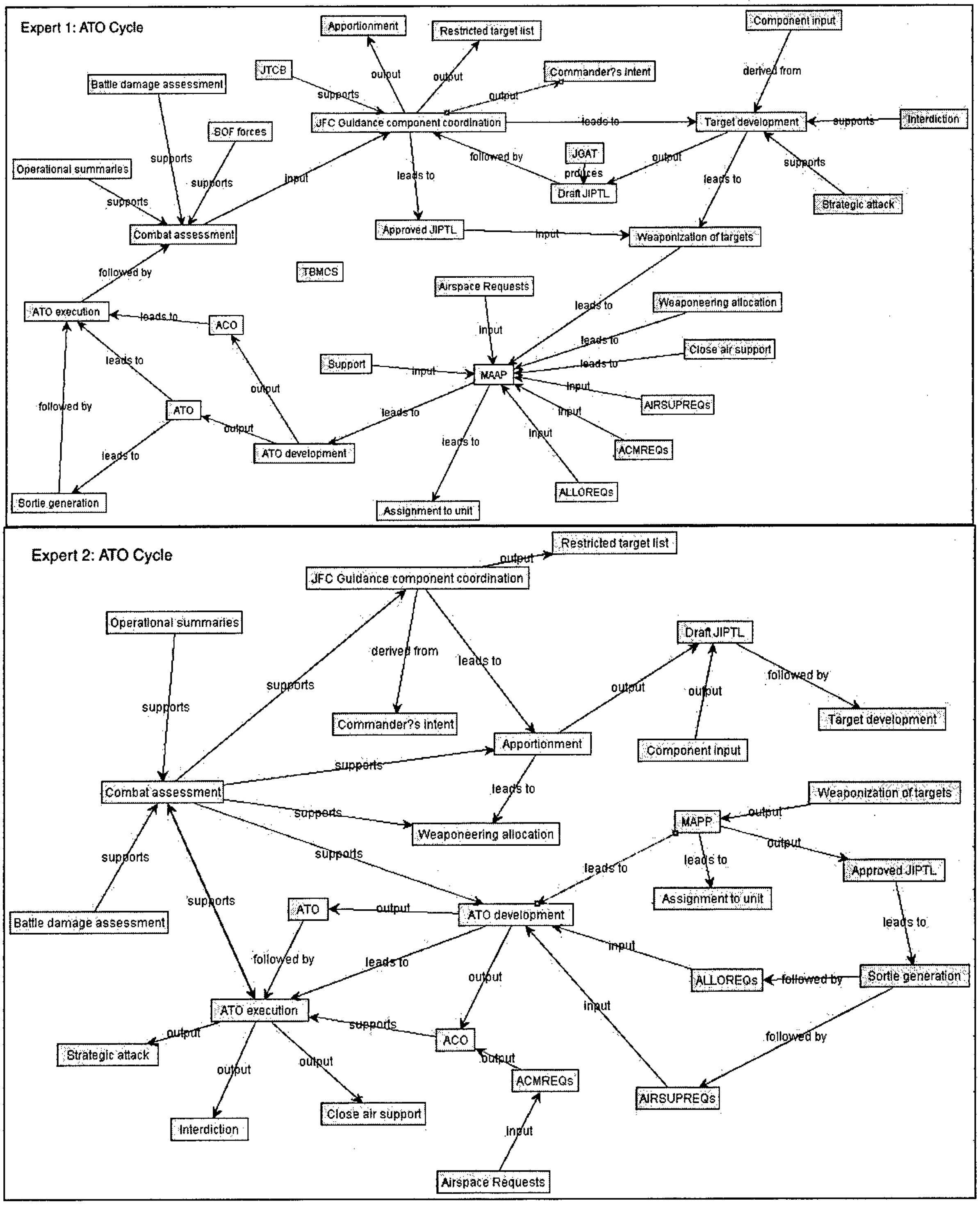

Figure 2. Expert knowledge maps for the ATO Cycle Task.

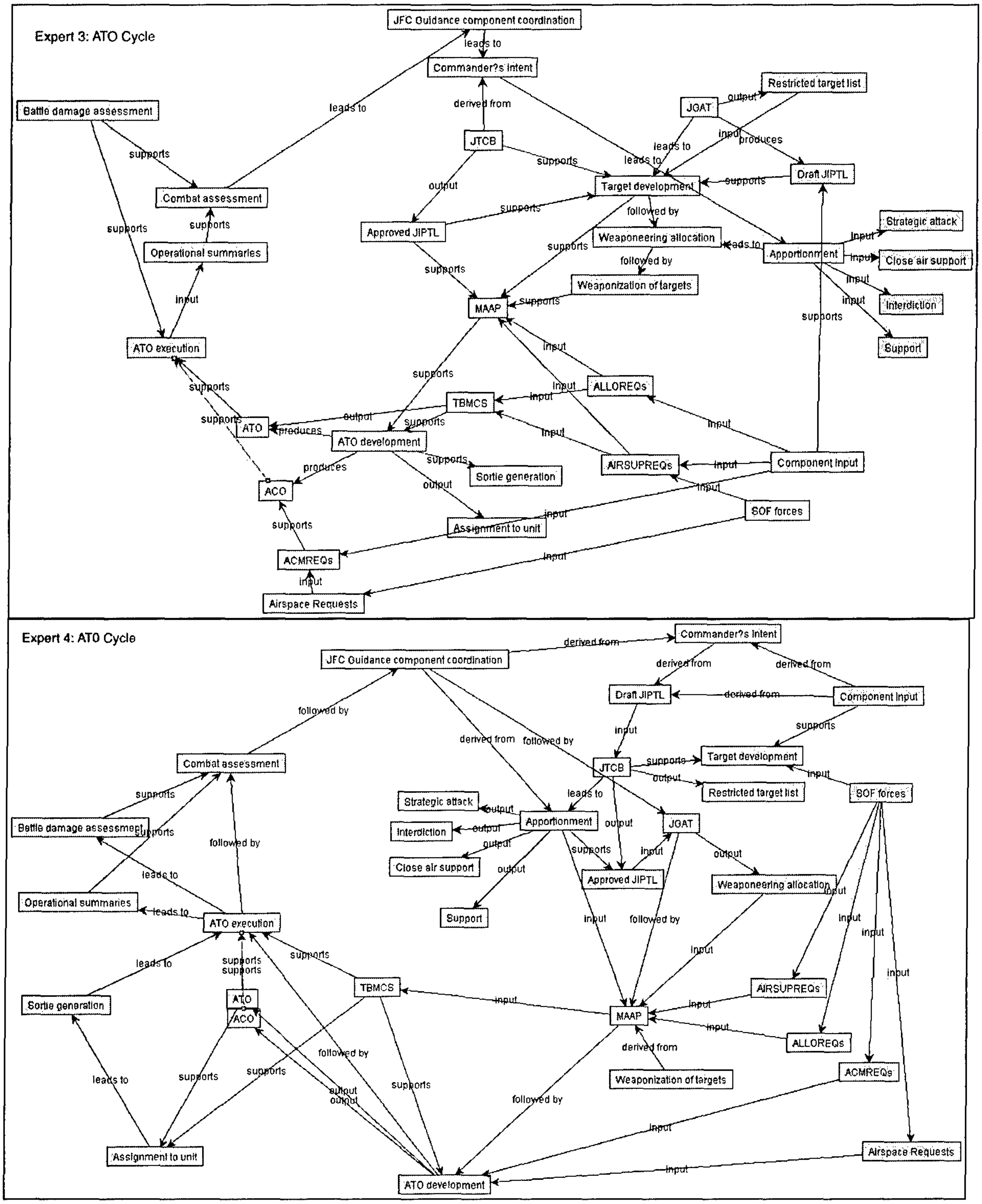

Figure 3. Expert knowledge maps for the ATO Cycle Task (continued).

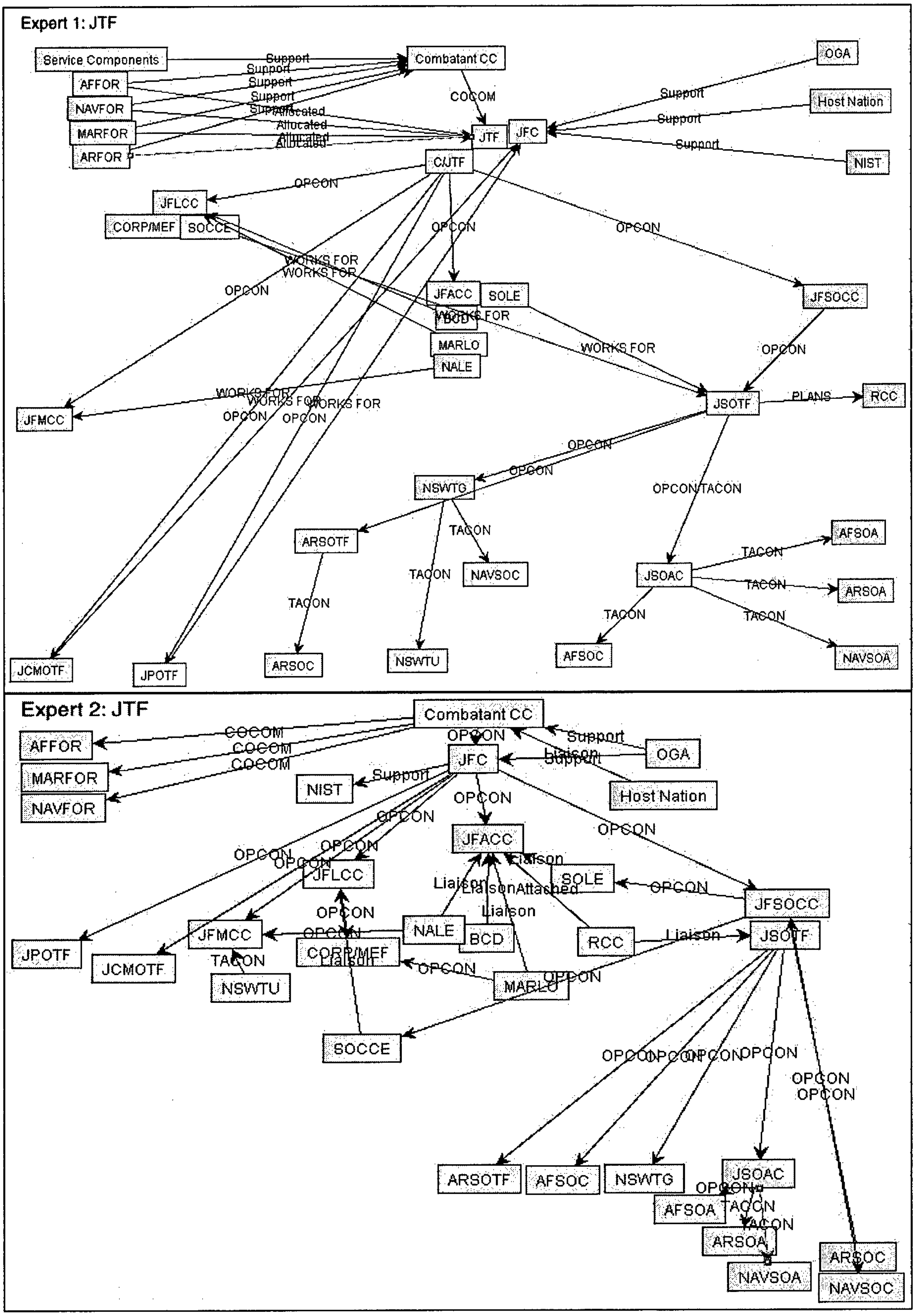

Figure 4. Expert knowledge maps for the (Joint Task Force) JTF task.

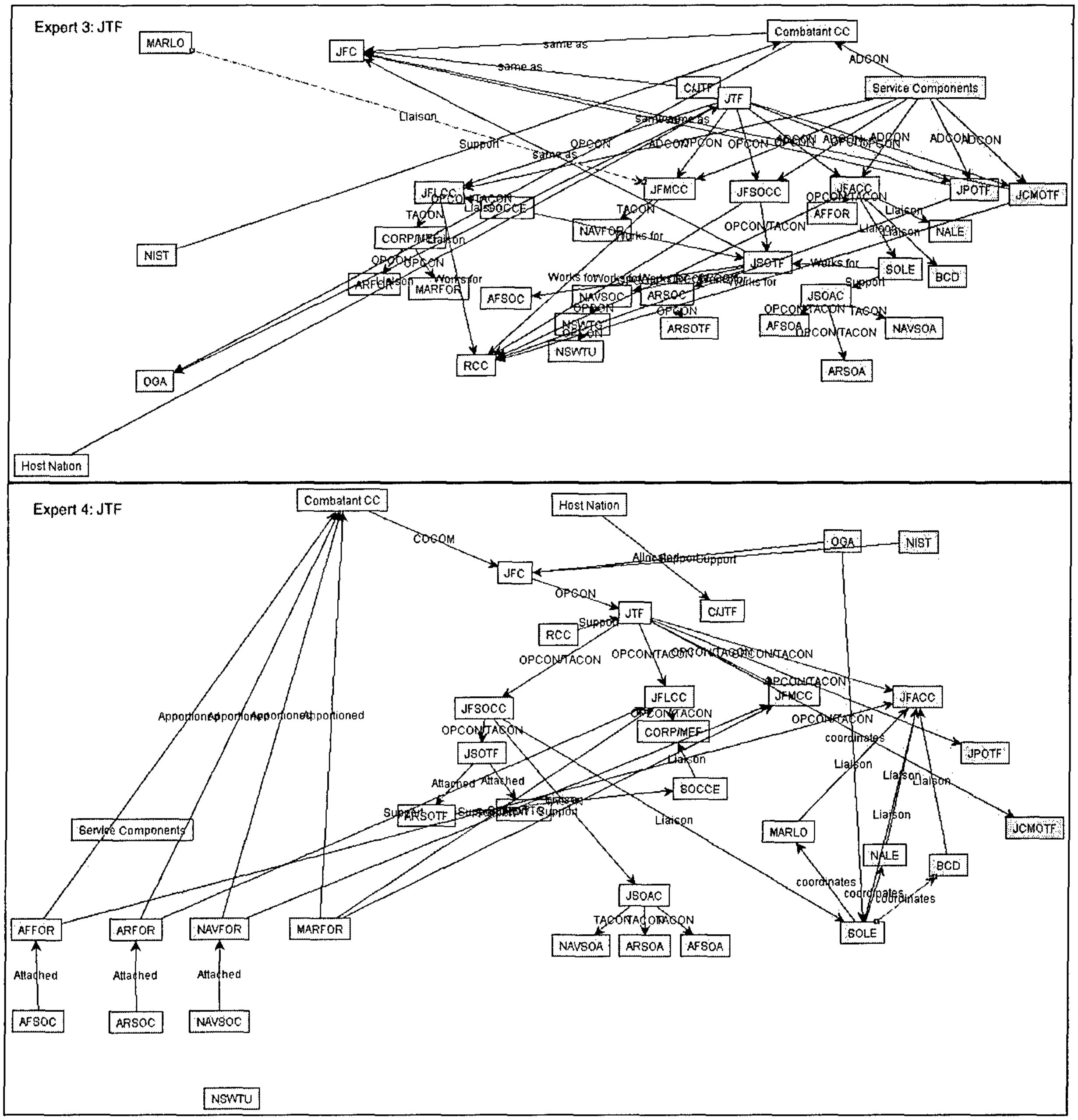

Figure 5. Expert knowledge maps for the (Joint Task Force) JTF task (continued).

Tasks and Measures

Participant knowledge map measures. Two mapping tasks, ATO and JTF, with predefined concepts and links were given to all participants after Day 1 of the course. Students were given 25 minutes to complete each map. Participants had access to the knowledge mapping tool until Day 4 to complete the third task, Joint Special Operations Task Force (JSOTF). Participants logged in to the Web site and launched the knowledge mapping tool, and were asked to save their maps onto our server at UCLA. Students worked individually without the aid of course material or notes. However, the instructor did provide them with a list of acronyms and abbreviations for some of the concepts in the mapping task.

Map Analysis

Preliminary grouping. A staff researcher took a preliminary look at the student maps and classified the students into four groups according to map density and organization across all three tasks. Density is defined as the number of terms and links used, as well as the number of clusters. Clusters are groups of related concepts gathered or occurring closely together around key ones. For example, in one expert map, <Apportionment> is a central idea around which concepts <Strategic attack>, <Interdiction>, <Close air support> and <Approved JIPTL> converge. The results are show in Table 7 by student ID.

Table 7

Preliminary Grouping of Students by Grade Performance on the Knowledge Maps

Grade	Student ID
Low	jsou101, 102, 107, 116, 120, 130
Medium/Low	jsou108, 113, 114, 117, 118, 119, 122, 124, 128
Medium/High	jsou103, 104, 109, 110, 115, 125, 126
High	jsou105, 106, 112, 121, 123, 127, 129

The analysis of the maps was reduced to two tasks, ATO and JTF, as future data collection opportunities were constrained to only two tasks due to time limitations.

Three methods of scoring. Following the preliminary grouping, three types of scoring methods were used to analyze the knowledge maps:

1. Automated criterion-based (expert) assessment
2. Propositional analysis
3. Structural mapping analysis

Automated criterion-based assessment. The first, automated criterion-based assessment is a scoring method based on the degree to which student maps contained the same or similar propositions (node-link-node) as compared to experts. The scores range from very stringent (or exact) matches to experts to very loose or less exact matching to the experts. The criterion-based scoring methods are summarized in Table 8.

Table 8

Criterion-based Scoring Methods for Knowledge Mapping Tasks

	Scoring scheme	Definition	Example
1	Exact Match	Number of propositions on student maps with an exact match in expert maps, taking into account both the direction of the relationship and the link label.	MAAP--inputs-->TBCMS
2	Directionless	Number of concept-concept matches, taking into account the link label between concepts but not the direction of the relationship.	MAAP--inputs--TBCMS
3	Linkless with Direction	Number of concept-concept matches, taking into account the direction of the relationship, but not the link label.	MAAP---->TBCMS
4	Linkless, Directionless	Number of concept-concept matches, not taking into account the direction of the relationship nor the link label.	MAAP----TBCMS

The stringency of scoring is highest for an Exact Match and lowest for Linkless, Directionless, with scores increasing accordingly. See, for example, the following ATO scores for jsou123 in Table 9.

Table 9

Sample Criterion-based Scores for One Participant on the ATO Cycle Knowledge Map

User Name	Exact	Directionless	Linkless, with Direction	Linkless, Directionless
jsou123	4	4	9	10

The experts performed significantly better than the students on all levels of scoring. The mean scores for students as scored based on experts and student experts are shown in Table 10.

Table 10

Mean Knowledge Mapping Scores of Students Compared to Experts by Task, Expert, and Scoring Type

ATO Cycle	Exact		Directionless		Linkless, with Direction		Linkless, Directionless		Expert
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Max Score
Expert1	0.84	0.85	1.04	0.89	3.6	2.3	5.16	2.5	35
Expert2	0.76	0.88	0.88	0.93	3.48	2.3	4.36	2.2	34
Expert3	0.8	1.08	1.2	1.2	3.84	2.2	6.4	2.7	47
Expert4	1.4	1.3	1.8	1.5	5.44	3.4	7.36	3.8	51
SE1	0.96	1	1.08	1.06	3.63	2.7	5.17	2.8	40
SE2	1.29	2.3	1.54	2.4	3.58	3.2	5.83	3.6	33
JTF	**Exact**		**Directionless**		**Linkless, with Direction**		**Linkless, Directionless**		**Expert**
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Max Score
Expert1	1.13	2.2	1.21	2.4	2.67	3.1	4.83	4.04	38
Expert2	1.58	2.2	1.75	2.4	3.92	4.1	5.13	5	36
Expert3	1	2	1.75	2.7	3.79	3.8	6.25	5.8	57
Expert4	1.42	2.1	1.46	2.2	3.83	3.9	5.67	4.8	51
SE1	1.43	2.4	2.22	3.2	3.35	4.4	6.43	6.1	44
SE2	0.91	1.6	2.35	2.7	3.09	3	5.91	5.3	47

Note. SE = student expert.

The table shows that students performed rather poorly, even with the most lenient scoring (linkless, directionless) compared to the maximum score (number of propositions in the expert map). Results are also shown for students' scores based on two different student experts. The experts, however, also disagreed with each other, showing that there is variability even among those considered to be experts in the content. This may be due to differences in interpretation of joint doctrine.

With respect to which scoring method yielded the best reliability, for the limited number of participants we had, the reliability was highest for the linkless with direction on the ATO Cycle task, and the highest for the method that does not take into account the link term or the direction of the link (linkless, directionless). Table 11 shows the alpha reliabilities for both tasks.

Table 11

Alpha Reliabilities for the ATO Cycle and JTF Tasks

Task	Exact	Directionless	Linkless, with Direction	Linkless, Directionless
ATO Cycle (n = 28)	.80	.75	.94	.92
JTF (n = 25)	.80	.85	.84	.87

Student-expert scoring differentiated levels of performance among students which reiterated classification by the researchers. Students classified in the *Low* group scored consistently lower with the student-expert scoring, while students classified in the *High* group scored consistently higher. There was more variability in scores with the students classified in *Low/Medium* and *Medium/High* groups. Interrater reliabilities for these classifications were significantly (.01 level) high [2 raters, Pearson's R = .96 for ATO Cycle (n = 18); Pearson's R = .91 for JTF (n = 16)].

Further analysis of student scores show mean student scores for the ATO Cycle were consistently higher with Expert 4 across all four scoring schemes (Exact; Directionless; Linkless with Direction; and Linkless, Directionless). Additionally, Expert 4 and Expert 3 ATO maps had higher correlations to each other than Expert 1 or Expert 2.

Mean student scores for the JTF task were generally higher than mean scores for the ATO task across all four scoring schemes. Additionally, mean student scores against all four experts were similar across all four scoring schemes.

The next step was to conduct a more detailed analysis correlating student background information (age, military branch/division, years in military service, years in Special Operations, highest level of education, pre- and post-instruction knowledge of tasks, and comfort level with computers) by task and across the two tasks. Student mean scores for the ATO cycle consistently higher with Expert 4 across all 4 scoring schemes. Expert 4 and Expert 3 correlated significantly higher with each other than with Expert 1 and 2. The ATO task correlated poorly against all combination of background measures. One possible explanation is that the ATO task is process-oriented, therefore allowing for greater variability in representation, whereas the JTF task is hierarchical/structural in representation. The mean student scores for JTF were generally higher than mean scores for ATO. Mean student scores against all four experts were comparable across all four scoring schemes.

In terms of background information, students with higher prior knowledge of JTF scored better on the JTF mapping task, and student self-report of learning more after having taken the course correlated with higher mapping scores. In addition, student comfort level with computers correlated positively with JTF scores. There were higher general correlations for JTF compared with ATO.

Propositional Analysis. The maps were also scored based on frequency of propositions across all students. Percentages of students using the propositions were calculated and put into a table and sorted from most frequent to least frequent. This was done both for the 4 experts and for the 29 students. The percentages between the two groups were compared and are reported in Table 12 and Table 13.

Table 12

Frequency of Source and Destination Terms and Links Used (ATO Cycle)

		Frequency (%) Expert	Frequency (%) Student
Source Terms	Apportionment	7.8%	**1.5%****
	ATO development	7.2%	6.4%
	JFC Guidance component coordination	7.2%	3.8%
	ATO execution	5.4%	6.8%
	Component input	5.4%	3.6%
	Combat assessment	4.8%	3.8%
	JTCB	4.8%	1.6%
	MAAP	4.8%	5.3%
	SOF Forces	4.8%	2.8%
Destination Terms	MAAP	12.0%	6.9%
	ATO execution	9.0%	4.6%
	Target development	7.8%	3.8%
	Combat assessment	7.2%	4.6%
	ATO development	6.6%	7.2%
	Draft JIPTL	4.8%	3.5%
Link Terms	supports	25.1%	14.5%
	input	21.6%	13.2%
	output	19.8%	3.3%
	leads to	18.0%	**32.5%****
	followed by	9.0%	14.3%

Table 13

Frequency of Source and Destination Terms and Links Used (JTF Structure and Function)

Type		Frequency (%) Expert	Frequency (%) Student
Source Terms	JTF	10.4%	8.4%
	JSOAC	8.2%	4.4%
	JSOTF	7.7%	10.6%
	JFSOCC	6.0%	4.1%
	Combatant CC	4.4%	6.0%
	JFC	4.4%	4.7%
	Service Components	4.4%	5.1%
	SOLE	4.4%	5.8%
	C/JTF	3.8%	2.5%
	SOCCE	3.8%	2.1%
	JFACC	8.8%	5.7%
	JFC	8.2%	**1.2%****
	JFLCC	7.7%	2.9%
	Combatant CC	7.1%	2.6%
	JFMCC	6.6%	2.1%
	JSOTF	5.5%	6.1%
	JTF	4.4%	4.0%
	JFSOCC	3.8%	2.5%
	RCC	3.8%	1.4%
Link Terms	OPCON	31.3%	17.3%
	TACON	16.5%	**1.8%****
	Liaison	11.5%	8.1%
	Support	11.5%	8.3%
	Works for	8.2%	13.0%
	ADCON	3.8%	0.8%

The expert and student use of terms and links were fairly proportional, with some exceptions. As can be seen in the tables, there were items where experts have different frequencies than those of students. For example, apportionment was the highest source term on the ATO cycle task for experts (at 7.8%) while for students it was much lower (1.5%). Experts relied on more functional links for ATO, e.g.

supports (25%), *input* (21%), *output* (19%), and students use more relational links for ATO, e.g. *leads to* (32%), *followed by* (14%), *supports* (14%).

Structural mapping analysis. The maps were also analyzed for their structure or interconnectedness. Structural mapping looks at the way knowledge is organized, in terms of a network of nodes. The focus is on how these nodes are connected to one another. The purpose of this type of analysis is to identify patterns in the knowledge space, and to identify how information flows through the knowledge system as a result of its structure. The components of the structural mapping analysis included:

- number of unique nodes
- nature of the node (source, sink or carrier): Source is a point of fan-outs (output) but not fan-ins (input), sink is a point of fan-ins (input) but not fan-outs (output), and carrier is a point of fan-ins (input) and fan-outs (output).
- number of fan-ins and fan-outs associated with each node Related to this clustering, in which one concept is the focal point for others. <MAAP> is considered a carrier whose links to <Weaponeering allocation>, <Close air support>, <AIRSUPREQs>, <ACMREQs>, and <ALLOREQs> constitute a cluster. Clustering is an important feature of map organization because it helps differentiate concepts into key concepts and supporting ones.
- clustering: groups of related concepts gathered
- reachability: defined as the accessibility of one concept to other concepts in the system, i.e., what other nodes are accessible to the node in question, providing information on connectedness

An analysis of the structure revealed differences between expert maps and student maps, and differences among students' maps relative to structural complexity. In general, the expert maps had more terms; variable use of source, sinks, and carriers; numerous clusters; and high reachability. Additionally, a comparison of a sample of student maps revealed similar patterns, with more sophisticated maps containing a higher number of terms, links, and clusters as well as level of reachability.

Among expert maps, the following key terms around which clusters occur were identified in Table 14.

Table 14

Clustering of Concepts Across Expert by Task

Task	Concepts (across experts)
ATO Cycle	Apportionment, ATO development, ATO execution, Combat assessment, Target development, and MAAP
JTF	Combatant CC, JSOTF, JFC, JSOTF, JFSOCC, JFACC, JFMCC, JFLCC and JSOAC

One of the goals of performing structural analyses and locating clusters is to identify areas of conceptual weakness. For example, if a map is missing one of the key terms above, or if the above key terms are missing or poorly elaborated by supporting terms, its paucity will provide important instructional and remediation feedback to instructors.

Generalizability. The generalizability analyses could not be performed due to an insufficient number of participants. We attempted a second round of data collection at Hurlburt, but technical difficulties related to communication between the JSOU lab computers and the UCLA server prevented access to the HPKMT. To avoid this problem in future data collection, we developed a standalone HPKMT that could collect data locally on each computer. Unfortunately, changes at JSOU ended their participation in the study, and ADL was not able to secure other sites for further data collection.

Discussion

Our findings have shown that knowledge mapping tasks like those used for assessment of the Joint Special Operations University (JSOU) courses elicited some valuable information regarding student understanding of course content. While there was some disagreement between experts, and between experts and students, the differences appeared to be greater on the ATO Cycle task, which is a more process-oriented map as compared to the JTF task which is a more hierarchical/structural map. The low performance of the students suggests that more remediation may be needed and/or more time given to participants to learn the content. It may also indicate that students need more exposure to this type of task, knowledge mapping, which in many cases could have been novel for them.

The questions regarding generalizability of the task remain. The results can be used not only for assessment purposes, but also for instructional remediation as determined by areas of weakness or misconceptions. Since the proposal of this study, a study conducted by other CRESST researchers, Yin and Shavelson (2004), with eighth-grade students in science (density and buoyancy) showed that reliability was greater with an S type mapping task (selecting link phrases) versus a C type (creating link phrases). In our study we did use the S type mapping task (predefined links). Some differences in method included scoring and whether links could be bidirectional. Their decision study found that there would need to be 18 to 20 mandatory propositions to get a G-coefficient near .80 from one occasion.

The three methods of scoring we used, criterion-based, propositional analysis, and structural analysis each yielded important information about the quality of the knowledge maps in relation to expert performance. The criterion-based method is currently automated, but the other two methods are not automatically scored, but could be. The scoring methods presented here should be further explored in future studies.

References

Advanced Distributed Learning. (2003a). *Advanced Distributed Learning Sharable Content Object Reference Model Version 1.2* [On-line]. Available: http://www.adlnet.org.

Advanced Distributed Learning. (2003b). *What works in distance learning*. H. F. O'Neil (Ed.). [On-line]. Available: http://www.adlnet.org.

Aguirre-Munoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners.* Unpublished doctoral dissertation. University of California, Los Angeles.

Air Force Institute for Advanced Distributed Learning (AFIADL). (2001). *Air Force Advanced Distributed Learning Vision*. Maxwell AFB-Gunter Annex, AL: Author.

Austin, L. B., & Shore, B. M. (1995). Using concept mapping for assessment in physics. *Physics Education, 30,* 41-45.

Cawley, J. M., Zimmaro, D. M., Van Meter, P., & Theodorou, E. S. (1999, April). *Validation of concept maps as a tool to predict performance on course exams*. Paper presentation at the annual meeting of the American Educational Research Association, Montreal, Canada.

Chung, G. K. W. K., Harmon, T. C., & Baker, E. L. (2001). The impact of a simulation-based learning design project on student learning. *IEEE Transactions on Education, 44,* 390–398.

Department of Defense (DoD). (1999). *The Department of Defense Advanced Distributed Learning Strategic Plan.* Washington, DC: Pentagon.

Director of Naval Training (N7). (1998). The *Navy-wide distributed learning planning strategy.* Washington, DC: Navy Pentagon.

Herl, H. E. (1995). *Construct validation of an approach to modeling cognitive structure of experts' and novices' U.S. history knowledge.* Unpublished doctoral dissertation, University of California, Los Angeles.

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of U.S. history knowledge. *Journal of Educational Research, 89,* 206-218.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15,* 315–334.

Joint Special Operations University. (2003). *Joint Special Operations University Course Handbook.* Hurlburt Field, FL: Author.

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, H. F., Jr. (2002). *The validity of knowledge mapping as a measure of elementary students' scientific understanding* (CSE Tech. Rep. No. 557). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lee, J. J. (1999). *The impact of Korean language accommodations on concept mapping tasks for Korean American English language learners.* Unpublished doctoral dissertation. University of California, Los Angeles.

Markham, K., Mintzes, J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching, 31,* 91-101.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36,* 475-492.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn.* New York: Cambridge University Press.

Osmundson, E., Chung, G. K. W. K., Herl, H. E., & Klein, D. C. D. (1999). *Concept mapping in the classroom: A tool for examining the development of students' conceptual understandings* (CSE Tech. Rep. No. 507). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Pearsall, N. R., Skipper, J., & Mintzes, J. (1997). Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Science Education, 81,* 193-215.

Rice, D., Ryan, J., & Samson, S. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching, 35,* 1103-1127.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33,* 569-600.

Ruiz-Primo, M. A., Schultz, S., Li, M., & Shavelson, R. J. (1997). *Concept map-based assessment in science: Two exploratory studies* (CSE Tech. Rep. No. 436). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38,* 260-278.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 15,* 403–418.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4,* 347–362.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

U.S. Army Training and Doctrine Command (TRADOC). (1999). *The Army Distance Learning Plan.* Fort Monroe, VA: Author

West, C. D., Pomeroy, J. R., Park, J. K., Gerstenberger, E. A., & Sandoval, J. (2000). Critical thinking in graduate medical education: A role for concept mapping assessment. *Journal of the American Medical Association, 284,* 1105-1110.

Yin, Y. & Shavelson, R.J. (2004). *Application of Generalizability Theory to concept-map assessment research.* (CSE Tech. Rep. No. 640). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Zimmaro, D. M., Zappe, S. M., Parkes, J. T., & Suen, H. K. (1999, April). *Validation of concept maps as a representation of structured knowledge.* Paper presentation at the annual meeting of the American Educational Research Association, Montreal, Canada.

Appendix: Abbreviations and Acronyms

A. Concepts

ACMREQs	Airspace Control Means Request
ACO	Airspace Control Order
AFFOR	Air Force Forces
AFSOA	Air Force Special Operations Aviation
AFSOC	Air Force Special Operations Command
AIRSUPREQs	Air Support Requests
ALLOREQs	Allocation Requests
ARFOR	Army Forces
ARSOA	Army Special Operations Aviation
ARSOC	Army Air Force Special Operations Command
ARSOTF	Army Special Operations Task Force
ATO	Air Tasking Order
BCD	Battlefield Coordination Detachment
C/JTF	Commander/ Joint Task Force
Combatant CC	Combatant Component Commander
CORP/MEF	CORP/Marine Expeditionary Force
Draft JIPTL	Joint Integrated Priority and Target List
JCMOTF	Joint Civil-military Operations Task Force
JFACC	Joint Force Air Component Commander
JFC	Joint Force Commander
JFLCC	Joint Force Land Component Commander
JFMCC	Joint Force Maritime Component Commander
JFSOCC	Joint Force Special Operations Component Commander
JGAT	Joint Guidance, Apportionment and Targeting
JPOTF	Joint Psychological Operations Task Force
JSOAC	Joint Special Operations Air Component
JSOTF	Joint Special Operations Task Force
JTCB	Joint Targeting Coordination Board
JTF	Joint Task Force
MAAP	Master Air Attack Plan
MARFOR	Marine Forces
MARLO	Marine Liaison Officer
NALE	Naval and Amphibious Liaison Element
NAVFOR	Navy Forces
NAVSOA	Naval Special Operations Aviation
NAVSOC	Naval Special Operations Component
NIST	national intelligence support team
NSWTG	Naval Special Warfare Task Group

NSWTU	Naval Special Warfare Task Unit
OGA	Other government agency
RCC	relocation coordination center; rescue coordination center
SOCCE	Special Operations Command and Control Element
SOLE	Special Operations Liaison Element
TBMCS	Theater Battle Management Core System

B. Links

ADCON	Administrative Control
COCOM	Combatant Command
OPCON	Operational Control
TACON	Tactical Control